The 3rd SFB 1102 PhD day

09:30-10:00	**Coffee**
10:00-10:30	**The Influence of Script Knowledge on On-Line Comprehension: Evidence from Event-Related Potentials** *Elisabeth Rabs (Project A1)*
10:30-11:00	**Integrating Words and Events in a Neural Network** *Clayton Greenberg (B4)*
11:00-11:30	**Learning to Generate - Inducing Sentence Planning Rules for Natural Language Generation** *Dave Howcroft (Project A4)*
11:30-12:00	**What can Shared Sequential Structures tell us about Linguistic Similarity?** *Andrea Fischer (Project C4)*
12:00-12:30	**Poster boasters**
12:30-13:15	**Lunch**
13:15-14:45	**Poster Session**
	Reading Polish with Czech Eyes: Mechanisms on Different Levels of a Written Intercomprehension Scenario *Klára Jágrová (Project C4)*
	Word Embeddings as Features for Supervised Coreference Resolution *Iliana Simova (Project B5)*
	Orthographic Intelligibility and Explanatory Variables in Slavic Reading Intercomprehension *Irina Stenger (Project C4)*
	Mel-cepstral Distortion of German Vowels in Different Information Density Contexts *Erika Brandt (Project C1)*
	The Interplay of Specificity and Referential Entropy Reduction in Situated Communication *Elli Tourtouri (Project C3)*
14:45-15:00	**Break**
15:00-16:00	**PhD meeting**

IDeaL SFB 1102

UNIVERSITÄT DES SAARLANDES

DFG

The Influence of Script Knowledge on On-Line Comprehension: Evidence from Event-Related Potentials

Elisabeth Rabs, Project A1

Psycholinguistic research has extensively studied the influence of linguistic context on language processing. It has been shown that words appearing in the context of (semantically) related words are easier to process, known as „lexical priming“. Less is known about the influence of non-linguistic context, e. g. script knowledge, defined as a person's knowledge about temporally and causally structured event sequences. In two ERP studies we investigate how script knowledge facilitates processing beyond what lexical priming alone can explain, and how it interacts with linguistic cues.

In Experiment 1 we use context sentences that either introduce a script or are neutral. Readers are then presented with either script-congruent or incongruent target nouns, after either minor ("a moment later") or major ("the next day") temporal shifts. These shifts are thought to be sensitive to script use, but not priming, since they mark the following event as happening within or outside of script boundaries. We find evidence for a decrease in N400 amplitude - known to reflect a word's unexpectedness or surprisal - for script-congruent nouns and minor temporal shifts, indicating that comprehenders are sensitive to temporal aspects of scripts, which cannot be explained by priming accounts.

In Experiment 2 (for an example see fig. 1) we present a context which introduces either one or two scripts, in the latter case one of them marked as inactive by an "instead-of"-construction. Readers are then presented with a target noun congruent with either the active or the inactive/unmentioned script. Again, we find a decrease in N400 amplitude for target nouns congruent with the active script vs. the inactive/unmentioned script. Additionally, we find a decreased N400 for inactive-script-congruent targets in a 2-script context, indicating that linguistic cues like "instead of" don't fully suppress the influence of the mentioned but inactive script. We therefore conclude that script knowledge activation interacts with linguistic cues, explaining N400 modulation in a way that priming accounts alone cannot.

Context Sentence:	
Introduction:	Roberta's cold had gotten worse.
2 Scripts:	Instead of going to the post office, she went to the pharmacy.
1 Script:	She went to the pharmacy.
Target Sentence:	
Active Script:	She entered and handed over the ***prescription*** with a smile.
Inactive Script:	She entered and handed over the ***package*** with a smile.

Figure 1. example sentence for Exp.2

Integrating Words and Events in a Neural Network

Clayton Greenberg, Project B4

In previous work, specifically the SRNN and LSRC architectures, we have explored how to incorporate large contexts in a neural network language model. But such large contexts contain well-studied linguistic structures such as event-participant and discourse relations. So, in this talk, we turn to the interpretability research questions: What kinds of information are we representing? Can we change the architecture to produce more "insightful" representations? Our vision for the first question is that the network should learn knowledge of event-level plausibility beyond simple commonality. For example, "croquet" is less common than "soccer", but has the same plausibility as an argument of "play". Then, for the second question, to push the network in this direction, we propose training for the joint objectives of word-level language modeling and thematic fit judgement estimation (multi-task learning). We will present a theoretical outline of the network architecture, and elicit some discussion on preprocessing of the WaCky corpus.

Learning to Generate - Inducing Sentence Planning Rules for Natural Language Generation

Dave Howcroft, Project A4

Off-the-shelf surface realization systems are relatively easy to find (e.g. OpenCCG, FUF/SURGE, SimpleNLG, inter alia), but developing NLG systems requires more than just turning trees into strings of words: we also have to choose what to talk about and *how* we want to talk about it. My work focuses on this question of *how*, given a high-level semantic representation, we should express the content of that representation.

I propose using Bayesian nonparametric methods to induce a mapping from chunks of semantic representation to chunks of the syntactic dependency representation used as input by OpenCCG. In my talk I'll introduce the corpus we've built to use as training data for this task, cover the model at a high level, and then provide some detail about our efforts to replicated related work on inducing tree substitution grammars.

What can Shared Sequential Structures tell us about Linguistic Similarity?

Andrea Fischer, Project C4

It is common practice to evaluate the similarity of languages by means of mathematical distances, i.e. pairwise measures based on comparing all pairs of the analyzed set of languages. Most often, these comparisons are based on the *regular correspondences* between cognate words in the languages. While such a pairwise approach is simple and straightforward to apply, it fails to capture the systematicity of correspondences between not only pairs, but sub-groups of the languages. In my talk, I will present a novel, information-theoretic approach which allows to do exactly this: assess linguistic similarity between all sub-groups of the analyzed languages. My approach leverages the joint compressibility of corresponding words from the given languages, and enables us to directly observe and interpret the degree of *sequential similarity* between the words' constructions. I will present preliminary results from the Slavic family, provide an overview of the current state of the project, and conclude with upcoming applications.

Reading Polish with Czech Eyes: Mechanisms on Different Levels of a Written Intercomprehension Scenario

Klára Jágrová, Project C4

The phenomenon of intercomprehension reveals a robust human ability to understand an unknown language, without being able to use it actively, i.e. for speaking of writing. This scenario works more or less well, depending on the language combination. Essentially, the relatedness of the perceived language to the native or further languages already available in the individual's linguistic repertoire is an indicator of how successful readers can be in deciphering the unknown language. The mutual intelligibility of a particular language combination can be described on different levels. This contribution aims at capturing the processes that take place when a Czech reader is trying to understand a Polish text.

The underlying hypothesis is that processing difficulty in intercomprehension results from factors on two orthogonal dimensions: (i) cross-lingual opacity (linguistic distance, or the opposite of *transparency*) and (ii) unpredictability in context (surprisal on the syntactic-semantic interface). Those parts that have a high linguistic distance and are relatively unpredictable are expected to cause the greatest difficulties to the reader.

On the dimension of opacity, there can be lexical, orthographic, and morphological divergences can result in a lack of understanding by the reader. In previous research (e.g. Herringa et al. 2013), it has been shown that corresponding distance measures on these levels correlate with human performance in intercomprehension. Lexical distance can be measured as the percentage of non-cognates (words that are etymologically not related to a translation in the reader's language and are therefore expected to be not understandable). Orthographic distance can be measured by means of Levenshtein distance or by rather sophisticated measures, such as conditional entropy and word adaptation surprisal. Also morphological distance can be measured with the same methods as orthographic distance.

On the dimension of context, a divergent linearization as well as the reader's semantic expectation may have a crucial influence on the successful disambiguation. In order to investigate the role of sentential context and to provide evidence of how "unexpected" or "unusual" a context can be, thinking-aloud-protocols with Czech native readers trying to decipher Polish stimuli were conducted. 32 Czech informants (16 pairs) were asked to cooperatively translate Polish sentences, which were displayed to them on two separate screens, into Czech. These informants were recorded during their work, so that not only written translation data, but also audio data about the deciphering process could be collected.

Surprisingly, even though individual words within the stimulus sentences were completely transparent or identical to Czech cognates, some informants entered wrong translations of the respective words. The audio recordings reveal that informants found these words not "fitting" into the respective sentences and therefore preferred a translation that would "make more sense". In an attempt to formally capture the unexpectedness of words in the stimuli (and the processing difficulty caused by this), the sentences were scored with the help of Kneser-Ney 3gram models trained on Czech. The surprisal scores of the stimuli sentences in the different linearization conditions correlate with the number of wrongly translated words in the experiments.

Word Embeddings as Features for Supervised Coreference Resolution

Iliana Simova, Project B5

A common reason for errors in coreference resolution is the lack of semantic information to help determine the compatibility between mentions referring to the same entity. Distributed representations, which have been shown successful in encoding relatedness between words, could potentially be a good source of such knowledge. Moreover, being obtained in an unsupervised manner, they could help address data sparsity issues in labeled training data at a small cost.

In this work we investigate whether and to what extend features derived from word embeddings can be successfully used for supervised coreference resolution. We experiment with several word embedding models, and several different types of embedding-based features, including embedding cluster and cosine similarity-based features. Our evaluations show improvements in the performance of a supervised state-of-the-art coreference system.

Orthographic Intelligibility and Explanatory Variables in Slavic Reading Intercomprehension

Irina Stenger, Project C4

Even though many studies in receptive multilingualism investigate the mutual intelligibility of written and spoken (closely) related languages, understanding of the contribution of orthography in reading intercomprehension is quite limited. In the present investigation, we present results of online free translation task experiments measuring the orthographic intelligibility of written East Slavic (Ukrainian, Belarusian) and South Slavic (Bulgarian, Macedonian, Serbian) languages obtained from Russian native speakers. We want to determine the role of orthography in written intercomprehension, and for that reason chose to focus on isolated cognate recognition first. The underlying assumption here is that the correct cognate recognition is the key to reading intercomprehension. If the reader correctly recognizes a minimal proportion of words, he or she will be able to piece together the meaning of the written message.

When investigating the role of orthography in Slavic intercomprehension the following question is crucial: How can orthographic intelligibility between related languages be predicted and explained? As explanatory variables we consider the effect of three computational methods: Levenshtein distance (LD), conditional entropy (CE) and word adaptation surprisal (WAS) and correlate these with intelligibility scores obtained in the online experiments. The results suggest that the mean normalized WAS and the full CE are better predictors of orthographic intelligibility than the mean normalized LD. Additionally, we correlate the results of the experiments with other factors that are likely to play a role in mutual intelligibility (Gooskens 2013, Kürschner et al. 2008, Vanhove and Berthele 2015): neighborhood density, word frequency, word length, and reading and translation time. The highest correlation was found in the negative correlation between word intelligibility and neighborhood density. Also reading and translation time as well as word length contribute significantly to orthographic intelligibility. However, the frequency of cognates in the reader's language is not a reliable predictor. This was already shown by Kürschner et al. (2008) and could be replicated in our experiments.

Focusing on orthographic intelligibility, orthographic correspondences themselves, as well as their frequency, their nature or their position can be expected to perform well as predictors of intelligibility. In future research, we will analyze mismatched orthographic correspondences more precisely in order to investigate what kind of correspondences either facilitate or hinder intercomprehension as well as to get qualitatively significant results.

Mel-cepstral Distortion of German Vowels in Different Information Density Contexts

Erika Brandt, Project C1

This study investigated whether German vowels differ significantly from each other in mel-cepstral distortion (MCD) when they stand in different information density (ID) contexts. We hypothesized that vowels in the same ID contexts are more similar to each other than vowels which stand indifferent ID conditions. Read speech material from PhonDat2 of 16 German natives (m = 10, f = 6) was analyzed. Bi-phone and word language models were calculated based on DeWaC. To account for additional variability in the data, prosodic factors, as well as corpus-specific frequency values were also entered into the statistical models. Results showed that vowels in different ID conditions were significantly different in their MCD values. Unigram word probability and corpus-specific word frequency showed the expected effect on vowel similarity with a hierarchy between non-contrasting and contrasting conditions. However, these did not form a homogeneous group since there were group-internal significant differences. The largest distance can be found between vowels produced at fast speech rate, and between unstressed vowels.

The Interplay of Specificity and Referential Entropy Reduction in Situated Communication

Elli Tourtouri, Project C3

In situated communication, reference can be established with expressions conveying either precise (Minimally-Specified, MS) or redundant (Over-Specified, OS) information. For example, while in Figure 1, "Find the blue ball" identifies exactly one object in all panels, only in the top displays is the adjective required. There is no consensus, however, concerning whether OS hinders processing (e.g., Engelhardt et al., 2011) or not (e.g., Tourtouri et al., 2015). Additionally, as incoming words incrementally restrict the referential domain, they contribute to the reduction of uncertainty regarding the target (i.e., referential entropy). Depending on the distribution of objects, the same utterance results in different entropy reduction profiles: "blue" reduces entropy by 1.58 bits in the right panels, and by .58 bits in the left ones, while "ball" reduces entropy by 1 and 2 bits, respectively. Thus, the adjective modulates the distribution of entropy reduction, resulting in uniform (UR) or non-uniform (NR) reduction profiles. This study seeks to establish whether referential processing is facilitated: a) by the use of redundant pre-nominal modification (OS), b) by the uniform reduction of entropy (cf. Jaeger, 2010), and c) when these two factors interact. Results from inspection probabilities and the Index of Cognitive Activity — a pupilometric measure of cognitive workload (Demberg & Sayeed, 2016) — indicate that processing was facilitated for both OS and UR, while fixation probabilities show a greater advantage for OS-UR. In conclusion, efficient processing is determined by both informativity of the reference and the rate of entropy reduction.

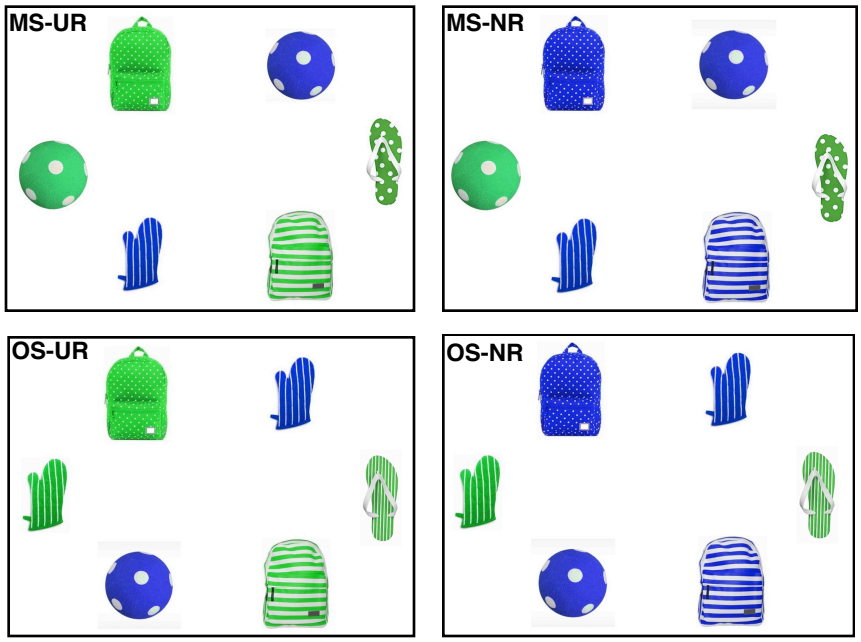

Figure 1. Sample visual stimuli, combined with the utterance "Find the blue ball".